The Canadian Journal of Statistics
Vol. 36, No. 3, 2008, Pages 443–461
La revue canadienne de statistique

Nonparametric adaptive likelihood weights

Jean-François PLANTE

Key words and phrases: Borrowing strength; Kuhn–Tucker conditions; likelihood method; nonparametric inference; nonparametric weights; weighted likelihood.

MSC 2000: Primary 62F10; secondary 62G05.

Abstract: The weighted likelihood can be used to make inference about one population when data from similar populations are available. The author shows heuristically that the weighted likelihood can be seen as a special case of the entropy maximization principle. This leads him to propose the minimum averaged mean squared error (MAMSE) weights. He describes an algorithm for calculating these weights and shows its convergence using the Kuhn–Tucker conditions. He explores the performance and properties of the weighted likelihood based on MAMSE weights through simulations.

Poids empiriques non paramétriques pour la vraisemblance

Résumé : La vraisemblance pondérée permet de faire de l'inférence sur une population en incorporant des données issues de populations semblables. L'auteur montre heuristiquement que la vraisemblance pondérée peut être vue comme un cas particulier du principe d'entropie maximale. Ceci le conduit à proposer les poids EQMIM (pour erreur quadratique moyenne intégrée minimale). Il décrit un algorithme pour le calcul de ces poids et en montre la convergence grâce aux conditions de Kuhn–Tucker. Il explore la performance et les propriétés de la vraisemblance pondérée basée sur les poids EQMIM à l'aide de simulations.

1. INTRODUCTION

The work of Stein (1956) showed that a biased estimate could sometimes be preferable to the best unbiased estimate as the biased one may compensate by featuring a smaller variance than the unbiased one. The modern terminology *borrowing strength* is most often used in a Bayesian setting, but refers in general to attempting to improve precision by using data from different sources. The weighted likelihood is designed to borrow strength while making minimal assumptions on the populations that are not of prime inferential interest.

The weighted likelihood that we study in this paper dates from the original work of Hu (1994), and its enhancements in Hu & Zidek (2002). The paradigm that we consider is a specific case of theirs that was introduced by Wang (2001) and further developed in Wang, van Eeden & Zidek (2004) and in Wang & Zidek (2005). We thus suppose that data comes from m distinct populations that have different yet similar distributions. More formally, for each fixed $i = 1, \ldots, m$,

$$X_{i1}, \ldots, X_{in_i} \overset{\text{iid}}{\sim} F_i$$

and we denote by f_i the corresponding density or mass function. Population 1 is of inferential interest, but the weighted likelihood

$$L_{\boldsymbol{\lambda}}(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(X_{ij} \,|\, \theta)^{\lambda_i / n_i}$$

lets other populations contribute to the inference so that the relevant information they contain is not lost. The density (or mass) functions $f(x \,|\, \theta)$ form a family of distributions indexed by $\theta \in \Theta$ such that there exists $\theta_0 \in \Theta$ for which $f(x \,|\, \theta_0) = f_1(x)$. The vector of exponents $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^{\mathsf{T}}$ downweights the contribution of the data according to the degree of similarity between the populations. The maximum weighted likelihood estimate (MWLE) is a value of θ that maximizes $L_{\boldsymbol{\lambda}}(\theta)$.

Ideally, the weights λ_i would be determined from scientific knowledge, but using data-based weights is more pragmatic. The ad hoc methods proposed by Hu & Zidek (2002) as well as the cross-validation weights of Wang (2001) and Wang & Zidek (2005) provide the only adaptive weights in the literature. None of these solutions is fully satisfactory.

The cross-validation weights, for instance, are flawed by instability problems. The simulation results in Wang (2001) and Wang & Zidek (2005) can be reproduced only at the cost of fine-tuning the algorithms numerically. In the ideal situation where some of the populations are identical to Population 1, the cross-validation weights may not even be defined.

Different methods allow one to borrow strength from other populations. Such methods typically rely on hierarchical models involving the $m-1$ populations that are not of prime inferential interest. Hierarchical models are primarily used in a Bayesian setting, which means that prior distributions for the hyperparameters must also be determined.

Efron (1996) develops an empirical Bayes method under a paradigm akin to ours where one population is deemed of prime interest. While he determines priors from the data in his work, the necessity of choosing a common model for all populations (in order to link them through hyperparameters) is not waived. By comparison, the weighted likelihood with the weights that we propose does not assume that the data follow a common model. The weights adapt, keeping the populations that are sufficiently similar to our target distribution, and dismissing the ones that are too different. The absence of parametric assumptions (except for the population of prime interest) may be a major advantage in situations where no natural or reliable hierarchical models are available, or when one is not completely comfortable with modeling the external populations. Being nonparametric, our weights are not subject to model misspecification.

It will become clear from the argument in Section 2 that the weighted likelihood is most useful when a mixture of the data from Population $2, \ldots, m$ is close to the target distribution. Such situations arise naturally in practice. Suppose, for instance, that previous studies based on gender, race, or other demographic variables, are available. Inference on the global population could be complemented by data from all the specific groups. When populations are likely to be similar, whether they come from adjacent geographical regions or akin populations, the weighted likelihood is also likely to produce good results since any mixtures of the populations will naturally be close to F_1.

In Section 2, we heuristically derive the weighted likelihood from Akaike's entropy maximization principle. That development justifies the formulation of the MAMSE weights that we define formally in Section 3. Section 4 presents some invariance properties of the MAMSE weights. An algorithm for computing these weights is proposed in Section 5: we use the Kuhn–Tucker sufficient conditions to show that it yields the desired solution. Simulation results appear in Section 6 where the performance of the MWLE with MAMSE weights is explored for different plausible scenarios. Finally, the asymptotic behavior of the MAMSE weights is briefly discussed in Section 7 and concluding remarks are in Section 8.

2. HEURISTIC JUSTIFICATION OF THE WEIGHTED LIKELIHOOD

Consider first the one-sample situation where n independent data points $Y_1, \ldots, Y_n$ come from a distribution whose unknown and unknowable density is $g(y)$. In his pioneering work, Akaike (1977) argues that the goal of inference should be the estimation of $g(y)$. When a parametric model $f(y \mid \theta)$ is to be used, Akaike proposes maximizing the relative entropy

$$B(g, f) = -\int \frac{g(y)}{f(y \mid \theta)} \log\left\{ \frac{g(y)}{f(y \mid \theta)} \right\} f(y \mid \theta)\, \mathrm{d}y.$$

The relative entropy is in fact minus the Kullback–Leibler divergence between f and g. In that case, it is a measure of the proximity of the distributions f and g. The expression for $B(g, f)$

can be further simplified

$$B(g,f) = -\int \log\left\{\frac{g(y)}{f(y\,|\,\theta)}\right\} g(y)\,\mathrm{d}y = \int \log\{f(y\,|\,\theta)\}g(y)\,\mathrm{d}y - \int \log\{g(y)\}g(y)\,\mathrm{d}y.$$

In particular, when the objective is to maximize $B(g,f)$ as a function of θ, the last term of the rightmost expression can be ignored since it does not depend on θ.

Calculating the entropy would require the knowledge of the unknown and unknowable true distribution g. We thus have to estimate it. Let

$$\widehat{G}(y) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(Y_i \le y)$$

be the empirical distribution function of the data set $Y_1, \ldots, Y_n$. The indicator variable $\mathbb{1}(\,\cdot\,)$ is equal to one if all the elements of its argument are true, and equal to 0 otherwise. Using $\mathrm{d}\widehat{G}(y)$ as an approximation to $\mathrm{d}G(y) = g(y)\,\mathrm{d}y$ yields

$$\int \log\{f(y\,|\,\theta)\}\,\mathrm{d}\widehat{G}(y) = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i\,|\,\theta),$$

the log-likelihood! Therefore, calculating the likelihood is equivalent to calculating the entropy where the true distribution is estimated by the empirical distribution of the data. Hence, the maximum likelihood estimate can be seen as a special case of Akaike's entropy maximization principle.

Consider now the m-population paradigm of Wang (2001) introduced earlier. With appropriate weights, the mixture $F_{\boldsymbol{\lambda}} = \sum_{i=1}^{m} \lambda_i F_i$ can be arbitrarily close to F_1. Let $\widehat{F}_i$ denote the empirical distribution function based on the sample from population i. The weighted empirical distribution function, written

$$\widehat{F}_{\boldsymbol{\lambda}} = \sum_{i=1}^{m} \lambda_i \widehat{F}_i \quad \text{with} \quad \lambda_i \ge 0 \quad \text{and} \quad \sum_{i=1}^{m} \lambda_i = 1,$$

called *relevance weighted empirical distribution* by Hu & Zidek (1993), may use more data than $\widehat{F}_1$, and thus be less variable. Hu & Zidek (1993) note the implicit bias involved in defacto replacing F_1 by $F_{\boldsymbol{\lambda}}$, but do not investigate as we do here the possibility of trading bias for precision.

In the context of maximum entropy, consider using the weighted empirical distribution function as an estimate of F_1. Then,

$$\int \log f(x\,|\,\theta)\,\mathrm{d}\widehat{F}_{\boldsymbol{\lambda}}(x) = \sum_{i=1}^{m} \lambda_i \int \log f(x\,|\,\theta)\,\mathrm{d}\widehat{F}_i(x) = \sum_{i=1}^{m} \frac{\lambda_i}{n_i} \sum_{j=1}^{n_i} \log f(X_{ij}\,|\,\theta),$$

the weighted log-likelihood! The maximum weighted likelihood can thus be derived from Akaike's entropy maximization principle.

Based on the heuristics above, a good choice of weights should try to make $\widehat{F}_{\boldsymbol{\lambda}}$ close to F_1. The criterion for determining weights presented in the next section is to try to achieve that goal. As a consequence, the weighted likelihood should be especially useful in situations where a mixture of the distributions of populations $2, \ldots, m$ is very similar to the target distribution. Such situations are not suitable for a hierarchical model which assumes that all data follow a common distribution whose parameters are allowed to differ from one group to another.

3. MINIMUM AVERAGED MEAN SQUARED ERROR WEIGHTS

Let us now turn to the central proposition of this paper, the definition of nonparametric adaptive likelihood weights.

As a prescreening step, samples that do not overlap with those of Population 1 are discarded by setting their weights to zero. For simplicity, the notation below does not reflect the possibly reduced number of populations considered: we suppose that m populations remain after prescreening.

The heuristics in Section 2 suggest that the maximum weighted likelihood estimate (MWLE) could perform well if $\widehat{F}_{\boldsymbol{\lambda}}$ were close to $\widehat{F}_1$ but would be less variable. These two requirements are combined in the objective function

$$P(\boldsymbol{\lambda}) = \int \left[\{\widehat{F}_1(x) - \widehat{F}_{\boldsymbol{\lambda}}(x)\}^2 + \widehat{\mathrm{var}}\{\widehat{F}_{\boldsymbol{\lambda}}(x)\}\right] \mathrm{d}\widehat{F}_1(x),$$

where the substitutions

$$\widehat{\mathrm{var}}\{\widehat{F}_{\boldsymbol{\lambda}}(x)\} = \sum_{i=1}^{m} \lambda_i^2 \, \widehat{\mathrm{var}}\{\widehat{F}_i(x)\} \quad \text{and} \quad \widehat{\mathrm{var}}\{\widehat{F}_i(x)\} = \frac{1}{n_i}\widehat{F}_i(x)\{1 - \widehat{F}_i(x)\}$$

are based on the distribution of the random variable $n_i\widehat{F}_i(x)$ that follows a binomial distribution for any fixed x.

We call the minimum averaged mean squared error (MAMSE) weights a vector of values $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^{\mathsf{T}}$ that solves the program

$$\begin{aligned} &\text{minimize } P(\boldsymbol{\lambda}) \\ &\text{subject to } \{\lambda_i \geq 0,\ i = 1, \ldots, m\} \text{ and } \sum_{i=1}^{m} \lambda_i = 1. \end{aligned}$$

The name MAMSE comes from the resemblance of the integrand with the mean squared error (Bias^2 + Variance).

4. STRUCTURAL PROPERTIES

Choosing the empirical distribution function to define the MAMSE weights implies some invariance properties that are discussed next.

THEOREM 1. *The minimum averaged mean squared error weights are invariant to a strictly increasing transformation of the data.*

Proof of Theorem 1. Let $X_{ij} = g(Y_{ij})$ where g is a strictly increasing function of the real line. Let H_i denote the cumulative distribution function of Y_{ij}. Then for all y, $x = g(y)$ and any $i = 1, \ldots, m$

$$\begin{aligned} \widehat{H}_i(y) &= \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbb{1}\{Y_{ij} \leq y\} = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbb{1}\{g(Y_{ij}) \leq g(y)\} \\ &= \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbb{1}\{X_{ij} \leq x\} = \widehat{F}_i(x). \end{aligned}$$

Since $P(\boldsymbol{\lambda})$ is integrated with respect to $\mathrm{d}\widehat{F}_1$, a discrete measure, there is no Jacobian of transformation in the integral and replacing all $\widehat{F}_i$ by the corresponding $\widehat{H}_i$ will not change the expression $P(\boldsymbol{\lambda})$, nor its maximum. □

THEOREM 2. *The minimum averaged mean squared error weights do not depend on the parametric model $f(x \mid \theta)$ used in the weighted likelihood.*

Proof of Theorem 2. The result follows immediately from the definition of the MAMSE weights and the choice of the nonparametric empirical distribution functions as estimates of F_i. □

THEOREM 3. *The maximum weighted likelihood estimate (MWLE) based on minimum averaged mean squared error (MAMSE) weights is invariant under a one-to-one reparameterization of $f(x \mid \theta)$ into $g(x \mid \tau) \triangleq f\{x \mid h(\tau)\}$, i.e., $\hat{\theta}$ is a MWLE iff $\hat{\tau}$ is a MWLE.*

Proof of Theorem 3. By Theorem 2, the MAMSE weights $\boldsymbol{\lambda}^* = [\lambda_1^*, \ldots, \lambda_m^*]^\mathsf{T}$ are invariant to the choice of parametric model $f(x \mid \theta)$. If τ is such that $\theta = h(\tau)$ and h is a one-to-one mapping of the parameter space, then $\tau_{\max}$ is such that

$$\prod_{i=1}^{m}\prod_{j=1}^{n_i} f\{X_{ij} \mid h(\tau)\}^{\lambda_i^*/n_i} \leq \prod_{i=1}^{m}\prod_{j=1}^{n_i} f\{X_{ij} \mid h(\tau_{\max})\}^{\lambda_i^*/n_i}$$

for all τ if and only if $\theta_{\max} = h(\tau_{\max})$ is such that

$$\prod_{i=1}^{m}\prod_{j=1}^{n_i} f(X_{ij} \mid \theta)^{\lambda_i^*/n_i} \leq \prod_{i=1}^{m}\prod_{j=1}^{n_i} f(X_{ij} \mid \theta_{\max})^{\lambda_i^*/n_i}$$

for all θ. Hence, the MWLE possesses the same functional invariance property as the maximum likelihood estimator (MLE) if we use the MAMSE weights. □

5. COMPUTING THE MAMSE WEIGHTS

Substituting $\lambda_1 = 1 - \sum_{i=2}^m \lambda_i$ allows one to embed the constraint $\sum_{i=1}^m \lambda_i = 1$ into the objective function $P(\boldsymbol{\lambda})$. Let us write

$$\tilde{\boldsymbol{\lambda}} = \begin{bmatrix} \lambda_2 \\ \vdots \\ \lambda_m \end{bmatrix}, \ \mathbf{V}(x) = \begin{bmatrix} \widehat{\mathrm{var}}\{\widehat{F}_2(x)\} & & 0 \\ & \ddots & \\ 0 & & \widehat{\mathrm{var}}\{\widehat{F}_m(x)\} \end{bmatrix}, \ \mathcal{F}(x) = \begin{bmatrix} \widehat{F}_1(x) - \widehat{F}_2(x) \\ \vdots \\ \widehat{F}_1(x) - \widehat{F}_m(x) \end{bmatrix}.$$

Then, the function $P(\boldsymbol{\lambda})$ can be written as

$$\begin{aligned} P(\boldsymbol{\lambda}) &= \int \Bigg[\Bigg\{ \widehat{F}_1(x) - \sum_{i=2}^m \lambda_i \widehat{F}_i(x) - \Bigg(1 - \sum_{i=2}^m \lambda_i\Bigg) \widehat{F}_1(x) \Bigg\}^2 \\ &\qquad + (1 - \tilde{\boldsymbol{\lambda}}^\mathsf{T}\mathbf{1})^2 \widehat{\mathrm{var}}\{\widehat{F}_1(x)\} + \sum_{i=2}^m \lambda_i^2 \widehat{\mathrm{var}}\{\widehat{F}_i(x)\} \Bigg] \mathrm{d}\widehat{F}_1(x) \\ &= \int \Big[\{\tilde{\boldsymbol{\lambda}}^\mathsf{T}\mathcal{F}(x)\}^2 + (1 - 2\tilde{\boldsymbol{\lambda}}^\mathsf{T}\mathbf{1} + \tilde{\boldsymbol{\lambda}}^\mathsf{T}\mathbf{1}\mathbf{1}^\mathsf{T}\tilde{\boldsymbol{\lambda}}) \widehat{\mathrm{var}}\{\widehat{F}_1(x)\} + \tilde{\boldsymbol{\lambda}}^\mathsf{T}\mathbf{V}(x)\tilde{\boldsymbol{\lambda}} \Big] \mathrm{d}\widehat{F}_1(x) \\ &= \int \Big[\tilde{\boldsymbol{\lambda}}^\mathsf{T} \big[\mathcal{F}(x)\mathcal{F}(x)^\mathsf{T} + \mathbf{V}(x) + \mathbf{1}\mathbf{1}^\mathsf{T} \widehat{\mathrm{var}}\{\widehat{F}_1(x)\} \big] \tilde{\boldsymbol{\lambda}} \\ &\qquad - 2\tilde{\boldsymbol{\lambda}}^\mathsf{T}\mathbf{1}\, \widehat{\mathrm{var}}\{\widehat{F}_1(x)\} + \widehat{\mathrm{var}}\{\widehat{F}_1(x)\} \Big] \mathrm{d}\widehat{F}_1(x) \\ &= \tilde{\boldsymbol{\lambda}}^\mathsf{T} \bar{\mathsf{A}} \tilde{\boldsymbol{\lambda}} - 2\tilde{\boldsymbol{\lambda}}^\mathsf{T}\mathbf{1}\bar{b} + \bar{b} \end{aligned} \tag{1}$$

where

$$\begin{aligned} \bar{\mathsf{A}} &= \int \big[\mathcal{F}(x)\mathcal{F}(x)^\mathsf{T} + \mathbf{V}(x) + \mathbf{1}\mathbf{1}^\mathsf{T} \widehat{\mathrm{var}}\{\widehat{F}_1(x)\} \big] \mathrm{d}\widehat{F}_1(x), \\ \bar{b} &= \int \widehat{\mathrm{var}}\{\widehat{F}_1(x)\} \mathrm{d}\widehat{F}_1(x). \end{aligned}$$

Hence, the minimum of $P(\boldsymbol{\lambda})$ without the constraints $\lambda_i \geq 0$ is the solution to the equation $\bar{\mathsf{A}}\tilde{\boldsymbol{\lambda}} = \bar{b}\mathbf{1}$. To ensure the weights are nonnegative, we apply the following algorithm and denote its solution by $\boldsymbol{\lambda}^*$ (or $\tilde{\boldsymbol{\lambda}}^*$).

1. Solve the equation $\bar{\mathsf{A}}\tilde{\boldsymbol{\lambda}} = \bar{b}\mathbf{1}$;
2. if all the weights obtained are nonnegative, stop. Otherwise set the negative weights to 0, ignore the corresponding samples and repeat from Step 1 with the reduced system. The weight allocated to Population 1 from Step 1 cannot be negative (see the proof of Lemma 4 for details). If no other samples are left, then $\tilde{\boldsymbol{\lambda}} = \mathbf{0}$ and $\lambda_1 = 1$.

The objective function $P(\boldsymbol{\lambda})$ is quadratic and positive (thus convex). Since the constraints form a convex set, intuition suggests that $\boldsymbol{\lambda}^*$ should be the global constrained minimum. Next we prove this more formally.

Consider the generic program

$$\begin{aligned} &\text{minimize } P(\boldsymbol{\lambda}) \\ &\text{subject to } \mathbf{h}(\boldsymbol{\lambda}) \leq \mathbf{0}, \end{aligned}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\mathbf{h}(\boldsymbol{\lambda}) = [h_1(\boldsymbol{\lambda}), \ldots, h_k(\boldsymbol{\lambda})]^\mathsf{T}$ is a vector of functions, each being from $\mathbb{R}^m$ to $\mathbb{R}$. Let $\boldsymbol{\nabla} P(\boldsymbol{\lambda})$ denote the gradient of P and $\mathbf{P}(\boldsymbol{\lambda})$ its Hessian. The same notation applies to $h_i(\boldsymbol{\lambda})$, $\boldsymbol{\nabla} h_i(\boldsymbol{\lambda})$ and $\mathbf{H}_i\ (\boldsymbol{\lambda})$. By definition, an $m \times m$ matrix B is positive definite (noted $B \succ 0$) if $\mathbf{y}^\mathsf{T} B \mathbf{y} > 0$ for all $\mathbf{y} \in \mathbb{R}^m \backslash \{\mathbf{0}\}$. The Kuhn–Tucker conditions (see, for instance, Luenberger 2003, p. 316) are as follows:

KUHN–TUCKER SECOND ORDER SUFFICIENCY CONDITIONS. *Let $h_1, \ldots, h_k$ and P be continuous and twice differentiable functions from $\mathbb{R}^m$ to $\mathbb{R}$. Sufficient conditions that a point $\boldsymbol{\lambda}^*$ be a strict relative minimum point of the program above are that there exists $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_k]^\mathsf{T} \in \mathbb{R}^k$ such that $\boldsymbol{\mu} \geq 0$, $\boldsymbol{\mu}^\mathsf{T} \mathbf{h}(\boldsymbol{\lambda}^*) = 0$,*

$$\boldsymbol{\nabla} P(\boldsymbol{\lambda}^*) + \sum_{i=1}^{k} \mu_i \boldsymbol{\nabla} h_i(\boldsymbol{\lambda}^*) = \mathbf{0} \tag{2}$$

and the matrix

$$\mathbf{P}(\boldsymbol{\lambda}^*) + \sum_{i=1}^{k} \mu_i \mathbf{H}_i(\boldsymbol{\lambda}^*) \succ 0. \tag{3}$$

Note that from Equation (1), we have $\boldsymbol{\nabla} P(\boldsymbol{\lambda}) = \bar{\mathsf{A}}\tilde{\boldsymbol{\lambda}} - \mathbf{1}\bar{b}$ and $\mathbf{P}(\boldsymbol{\lambda}) = \bar{\mathsf{A}}$. The function $P(\boldsymbol{\lambda})$ and its derivatives do not depend on λ_1 since it was replaced by $\lambda_1 = 1 - \mathbf{1}^\mathsf{T}\tilde{\boldsymbol{\lambda}}$. Consequently, it is implicitly understood in the following that P and h_i are functions of $\tilde{\boldsymbol{\lambda}} \in \mathbb{R}^{m-1}$, even when we write $P(\boldsymbol{\lambda})$ and $h_i(\boldsymbol{\lambda})$ rather than $P(\tilde{\boldsymbol{\lambda}})$ and $h_i(\tilde{\boldsymbol{\lambda}})$.

LEMMA 1. *The Hessian matrix $\mathbf{P}(\boldsymbol{\lambda})$ is positive definite.*

Proof of Lemma 1. Remember that

$$\bar{\mathsf{A}} = \int \left[\mathcal{F}(x)\mathcal{F}(x)^\mathsf{T} + \mathbf{V}(x) + \mathbf{1}\mathbf{1}^\mathsf{T} \widehat{\mathrm{var}}\{\widehat{F}_1(x)\}\right] \mathrm{d}\widehat{F}_1(x).$$

For any fixed x, each term of the integrand as written above is nonnegative definite. In particular, for any $\mathbf{y} \in \mathbb{R}^{m-1} \backslash \{\mathbf{0}\}$,

$$\mathbf{y}^\mathsf{T}\left\{\mathcal{F}(x)\mathcal{F}(x)^\mathsf{T} + \mathbf{1}\mathbf{1}^\mathsf{T} \widehat{\mathrm{var}}\{\widehat{F}_1(x)\}\right\}\mathbf{y} \geq 0.$$

The prescreening that we applied before optimizing $P(\boldsymbol{\lambda})$ ensures that all remaining samples overlap with that of Population 1. Therefore, $\widehat{\mathrm{var}}\{\widehat{F}_i(x)\} > 0$ for at least one of the X_{1j}, $j = 1, \ldots, n_1$ and thus

$$\int \widehat{\mathrm{var}}\{\widehat{F}_i(x)\}\, \mathrm{d}\widehat{F}_1(x) > 0.$$

Consequently, the diagonal elements of $\int \mathbf{V}(x)\, \mathrm{d}\widehat{F}_1(x)$ are strictly positive, which means that

$$\mathbf{y}^{\mathsf{T}}\left[\int \mathbf{V}(x)\, \mathrm{d}\widehat{F}_1(x)\right]\mathbf{y} = \sum_{i=1}^{m-1} y_i^2\left[\int \widehat{\mathrm{var}}\{\widehat{F}_i(x)\}\, \mathrm{d}\widehat{F}_1(x)\right] > 0$$

for any $\mathbf{y} \in \mathbb{R}^{m-1}\backslash\{\mathbf{0}\}$. Therefore, $\mathbf{y}^{\mathsf{T}}\bar{\mathsf{A}}\mathbf{y} > 0$, i.e., the Hessian of P, $\mathbf{P}(\boldsymbol{\lambda}) = \bar{\mathsf{A}}$, is positive definite. □

COROLLARY 1. *Equation* (3) *is satisfied.*

Proof of Corollary 1. In our implementation of the general Kuhn–Tucker conditions, $h_i(\boldsymbol{\lambda}) \equiv -\lambda_{i+1}$. Therefore, $\mathbf{H}_i(\boldsymbol{\lambda}) \triangleq \boldsymbol{\nabla}^{\mathsf{T}}\boldsymbol{\nabla}h_i(\boldsymbol{\lambda}) = \mathbf{0}$ are null matrices. From Lemma 1, we know that $\mathbf{P}(\boldsymbol{\lambda}^*)$ is positive definite, hence Equation (3) is satisfied. □

Applying the algorithm above will change negative weights to $\lambda_{i+1} = 0$ for some $i \in I^C \subset \{1, \ldots, m-1\}$ where I^C may be null. The set I contains the remaining indices and may also be null.

Let $J \subset \{1, \ldots, m-1\}$ be a possibly null subset of indices, then $A_{I,J}$ is the submatrix of A for the rows $i \in I$ and the columns $j \in J$. We define the subvector $\boldsymbol{\lambda}_I$ similarly.

The proposed algorithm involves solving reduced systems where the rows and columns for $i \in I^C$ are excluded. The system of equations that has to be solved then involves the matrix

$$\mathcal{A}_I = \int \left[\mathcal{F}_I(x)\mathcal{F}_I(x)^{\mathsf{T}} + \mathbf{V}_{I,I}(x) + \mathbf{1}_I\mathbf{1}_I^{\mathsf{T}}\widehat{\mathrm{var}}\{\widehat{F}_1(x)\}\right] \mathrm{d}\widehat{F}_1(x).$$

For convenience of exposition, suppose that the order of appearance of the λ_i in $\tilde{\boldsymbol{\lambda}}$ is such that all the λ_i that are "forced" to be zero are last. Then, with

$$\mathcal{F}(x) = \left[\begin{array}{c} \mathcal{F}_I(x) \\ \hdashline \mathcal{F}_{I^C}(x) \end{array}\right],$$

we can write

$$\begin{aligned}
\bar{\mathsf{A}} &= \int \left[\left[\begin{array}{c} \mathcal{F}_I(x) \\ \hdashline \mathcal{F}_{I^C}(x) \end{array}\right]\left[\begin{array}{c} \mathcal{F}_I(x) \\ \hdashline \mathcal{F}_{I^C}(x) \end{array}\right]^{\mathsf{T}} + \mathbf{V}(x) + \mathbf{1}\mathbf{1}^{\mathsf{T}}\widehat{\mathrm{var}}\{\widehat{F}_1(x)\}\right] \mathrm{d}\widehat{F}_1(x) \\
&= \int \left[\begin{array}{c:c} \mathcal{F}_I(x)\mathcal{F}_I(x)^{\mathsf{T}} + \mathbf{V}_{I,I}(x) & \mathcal{F}_I(x)\mathcal{F}_{I^C}(x)^{\mathsf{T}} \\ \hdashline \mathcal{F}_{I^C}(x)\mathcal{F}_I(x)^{\mathsf{T}} & \mathcal{F}_{I^C}(x)\mathcal{F}_{I^C}(x)^{\mathsf{T}} + \mathbf{V}_{I^C,I^C}(x) \end{array}\right] \\
&\quad + \mathbf{1}\mathbf{1}^{\mathsf{T}}\widehat{\mathrm{var}}\{\widehat{F}_1(x)\}\, \mathrm{d}\widehat{F}_1(x) \\
&= \left[\begin{array}{c:c} \bar{\mathsf{A}}_{I,I} & \bar{\mathsf{A}}_{I,I^C} \\ \hdashline \bar{\mathsf{A}}_{I^C,I} & \bar{\mathsf{A}}_{I^C,I^C} \end{array}\right].
\end{aligned}$$

In particular, $\mathcal{A}_I = \bar{\mathsf{A}}_{I,I}$ and $\mathcal{A}_{I^C} = \bar{\mathsf{A}}_{I^C,I^C}$. Therefore, the last step of the proposed algorithm is to solve the system of equations $\mathcal{A}_I\tilde{\boldsymbol{\lambda}}_I = \bar{\mathsf{A}}_{I,I}\tilde{\boldsymbol{\lambda}}_I = \mathbf{1}_I\bar{b}$. Note that this implies that the matrix $\mathcal{A}_I$ need not be recalculated at each step of the algorithm.

LEMMA 2. *If $I^C \neq \emptyset$ and $\mathbf{y} \in \mathbb{R}^{m-1} \backslash \{\mathbf{0}\}$ is any nonnegative vector with $\mathbf{y}_I = \mathbf{0}$ and $\mathbf{y}_{I^C} > \mathbf{0}$, then $\boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)^\mathsf{T}\mathbf{y} > \mathbf{0}$.*

Proof of Lemma 2. First note that the expression $\boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)^\mathsf{T}\mathbf{y}$ corresponds to the directional derivative of P at $\boldsymbol{\lambda}^*$ in the direction $\mathbf{y}$. Next consider the unit vector $\mathbf{e}_i \in \mathbb{R}^{m-1}$ whose i^{th} element is 1. For $i \in I^C$, the global unconstrained minimum of the convex function P is outside of the half-space $\lambda_{i+1} \geq 0$. Therefore, P increases in the direction $\mathbf{e}_i$ at $\boldsymbol{\lambda}^*$ and thus $\boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)^\mathsf{T}\mathbf{e}_i > 0$.

Finally, the hypothesized vector $\mathbf{y}$ can be expressed as a linear combination of vectors $\{\mathbf{e}_i : i \in I^C\}$ with nonnegative coefficients y_i. Therefore,

$$\boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)^\mathsf{T}\mathbf{y} = \sum_{i \in I^C} y_i \boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)^\mathsf{T}\mathbf{e}_i > 0. \qquad \square$$

Although $I = \emptyset$ or $I^C = \emptyset$ may occur, the following proofs hold under these special cases.

LEMMA 3. *The proposed algorithm solves the quadratic program*

$$\begin{aligned} &\textit{minimize } P(\boldsymbol{\lambda}) \\ &\textit{subject to } \{\lambda_i \geq 0, i = 2, \ldots, m\}. \end{aligned}$$

Proof of Lemma 3. To verify that the Kuhn–Tucker conditions are satisfied, first note that for $i = 1, \ldots, m-1$ the functions $h_i(\boldsymbol{\lambda}) \equiv -\lambda_{i+1}$ are continuous and twice differentiable. The quadratic objective function $P(\boldsymbol{\lambda})$ shares the same smoothness properties. Moreover, Corollary 1 establishes that Equation (3) holds.

At termination, the algorithm yields $\tilde{\boldsymbol{\lambda}}_I^* \geq \mathbf{0}$ and $\tilde{\boldsymbol{\lambda}}_{I^C}^* = \mathbf{0}$. The proposed solution $\boldsymbol{\lambda}^*$ is thus in the feasible set. It remains to show that there exists a $\boldsymbol{\mu}$ that satisfies the Kuhn–Tucker conditions stated earlier. We will show that $\boldsymbol{\mu} = \boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)$ satisfies the required properties. Expression (2) can be written

$$\boldsymbol{\nabla} P(\boldsymbol{\lambda}^*) + \sum_{i=1}^{m-1} \mu_i(-\mathbf{e}_i) = \boldsymbol{\nabla} P(\boldsymbol{\lambda}^*) - \boldsymbol{\mu} = \mathbf{0}$$

and clearly holds for $\boldsymbol{\mu} = \boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)$. The other Kuhn–Tucker conditions require that $\boldsymbol{\mu} \geq \mathbf{0}$ and $\boldsymbol{\mu}^\mathsf{T}\tilde{\boldsymbol{\lambda}}^* = \mathbf{0}$.

$\underline{\boldsymbol{\mu} \geq \mathbf{0}.}$

The last step of the algorithm before termination is to solve $\bar{\mathsf{A}}_{I,I}\tilde{\boldsymbol{\lambda}}_I = \mathbf{1}_I \bar{b}$. Therefore,

$$\boldsymbol{\mu}_I = \boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)_I = \left[\bar{\mathsf{A}}\tilde{\boldsymbol{\lambda}}^* - \mathbf{1}\bar{b}\right]_I = \bar{\mathsf{A}}_{I,I}\tilde{\boldsymbol{\lambda}}_I^* + \bar{\mathsf{A}}_{I,I^C}\tilde{\boldsymbol{\lambda}}_{I^C}^* - \mathbf{1}_I\bar{b} = \mathbf{0}$$

since $\tilde{\boldsymbol{\lambda}}_{I^C}^* = \mathbf{0}$.

In addition, we have from Lemma 2 that $\mu_i = \boldsymbol{\mu}^\mathsf{T}\mathbf{e}_i = \boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)\mathbf{e}_i > 0$ for all $i \in I^C$, and hence $\boldsymbol{\mu}_{I^C} > \mathbf{0}$. Therefore, $\boldsymbol{\mu} \geq \mathbf{0}$.

$\underline{\boldsymbol{\mu}^\mathsf{T}\tilde{\boldsymbol{\lambda}}^* = \mathbf{0}.}$

We can write the condition $\boldsymbol{\mu}^\mathsf{T}\tilde{\boldsymbol{\lambda}}^* = 0$ as $\boldsymbol{\mu}_I^\mathsf{T}\tilde{\boldsymbol{\lambda}}_I^* + \boldsymbol{\mu}_{I^C}^\mathsf{T}\tilde{\boldsymbol{\lambda}}_{I^C}^* = 0$. It is shown above that $\boldsymbol{\mu}_I = \mathbf{0}$, hence $\boldsymbol{\mu}_I^\mathsf{T}\tilde{\boldsymbol{\lambda}}_I^* = 0$. Moreover, the definition of the set I implies that $\tilde{\boldsymbol{\lambda}}_{I^C}^* = \mathbf{0}$, thus $\boldsymbol{\mu}_{I^C}^\mathsf{T}\tilde{\boldsymbol{\lambda}}_{I^C}^* = 0$ and the condition is satisfied.

Consequently, the solution found by the proposed algorithm is a strict relative minimum since it satisfies the sufficient Kuhn–Tucker conditions. $\square$

LEMMA 4. *The solution found by the proposed algorithm also satisfies the additional constraint* $\sum_{i=2}^{m} \lambda_i < 1$*, or equivalently,* $\lambda_1 > 0$.

Proof of Lemma 4. The solution found by the algorithm satisfies $\bar{\mathsf{A}}_{I,I}\boldsymbol{\lambda}_I^* = \mathbf{1}_I\bar{b}$. Expanding $\bar{\mathsf{A}}_{I,I}$ in this equation yields

$$\left[\int \left[\mathcal{F}_I(x)\mathcal{F}_I(x)^{\mathsf{T}} + \mathbf{V}_{I,I}(x) + \mathbf{1}_I\mathbf{1}_I^{\mathsf{T}}\widehat{\mathrm{var}}\{\widehat{F}_1(x)\}\right] \mathrm{d}\widehat{F}_1(x)\right]\tilde{\boldsymbol{\lambda}}_I^* = \mathbf{1}_I\bar{b}.$$

By subtracting $\bar{b}\mathbf{1}_I\mathbf{1}_I^{\mathsf{T}}\tilde{\boldsymbol{\lambda}}_I^*$ from both sides and multiplying the resulting equation by $\tilde{\boldsymbol{\lambda}}_I^{*\mathsf{T}}$ on the left, we have

$$\begin{aligned}\tilde{\boldsymbol{\lambda}}_I^{*\mathsf{T}}\left[\int \left\{\mathcal{F}_I(x)\mathcal{F}_I(x)^{\mathsf{T}} + \mathbf{V}_{I,I}(x)\right\} \mathrm{d}\widehat{F}_1(x)\right]\tilde{\boldsymbol{\lambda}}_I^* &= \bar{b}\tilde{\boldsymbol{\lambda}}_I^{*\mathsf{T}}(\mathbf{1}_I - \mathbf{1}_I\mathbf{1}_I^{\mathsf{T}}\tilde{\boldsymbol{\lambda}}_I^*) \\ &= \bar{b}\tilde{\boldsymbol{\lambda}}_I^{*\mathsf{T}}\mathbf{1}_I(1 - \mathbf{1}_I^{\mathsf{T}}\tilde{\boldsymbol{\lambda}}_I^*) = \bar{b}\left(1 - \sum_{i\in I}\lambda_{i+1}^*\right)\left(\sum_{i\in I}\lambda_{i+1}^*\right).\end{aligned}$$

By the same argument as in the proof of Lemma 1, the matrix on the left hand-side is positive definite, and hence the expression itself is positive. Since $\bar{b}$ and $\tilde{\boldsymbol{\lambda}}_I^*$ are positive, we necessarily have $1 - \sum_{i\in I}\lambda_{i+1}^* > 0$. Hence, the solution to the program in Lemma 3 always satisfies the additional constraint $\sum_{i\in I}\lambda_{i+1}^* = \sum_{i=2}^{m}\lambda_i^* < 1$ (remember that $\tilde{\boldsymbol{\lambda}}_{I^C}^* = \mathbf{0}$). This inequality is equivalent to $\lambda_1^* > 0$.

Regarding the comment to the effect that λ_1 cannot be negative for intermediate steps, consider the development above for such steps where $\boldsymbol{\lambda}_I$ may still contain negative values. Note that the left-hand side of the expression is still positive because of its positive definiteness. Moreover, the right-hand side can be written as $\lambda_1(1-\lambda_1)\bar{b}$, which means that $\lambda_1(1-\lambda_1)$ is positive. Therefore, $\lambda_1 \in (0,1)$, except if $I = \emptyset$ in which case $\lambda_1 = 1$ and $\tilde{\boldsymbol{\lambda}} = \mathbf{0}$. □

THEOREM 4. *The proposed algorithm solves the quadratic program*

$$\begin{aligned}&\textit{minimize } P(\boldsymbol{\lambda}) \\ &\textit{subject to } \{\lambda_i \geq 0,\ i = 1,\ldots,m\} \textit{ and } \sum_{i=1}^{m}\lambda_i = 1.\end{aligned}$$

Proof of Theorem 4. The result follows from Lemmas 3 and 4. □

6. SIMULATIONS

In this section, the finite-sample performance of the MWLE with MAMSE weights is evaluated through simulations. Different cases of interest are considered.

The number of repetitions for each simulation study varies from 10000 to 40000. We used the bootstrap on a pilot simulation to evaluate the variability of the values presented throughout this section. Unless otherwise stated, the standard deviation of the error due to simulation is less than one unit of the last digit shown.

6.1. Two normal distributions.

We first explore the merits of our weights for the ubiquitous normal distribution. Samples of equal sizes n are drawn from

$$\text{Pop. } 1 : \mathsf{N}(0,1), \quad \text{Pop. } 2 : \mathsf{N}(\Delta,1)$$

for different values of Δ, each scenario being repeated 10000 times. Table 1 shows the average MAMSE weights under different circumstances.

TABLE 1: Average MAMSE weights for Population 1 when equal samples of size n are drawn from normal distributions with unit variance and means 0 and Δ respectively. The results are averages over 10000 replicates.

	Average values of $100\lambda_1$									
	$n = 5$	10	15	20	25	50	100	200	1000	10000
$\Delta = 0$	72	71	72	71	71	72	72	72	72	72
0.001	72	71	71	72	72	72	72	71	72	72
0.01	72	72	71	72	72	72	72	72	72	74
0.10	72	72	73	73	73	73	74	76	86	98
0.25	74	74	75	76	76	79	83	88	97	100
0.50	77	79	80	82	83	88	93	96	99	100
0.75	80	83	86	88	89	94	97	98	100	100
1.00	84	87	90	92	93	96	98	99	100	100
1.50	89	92	94	95	96	98	99	99	100	100
2.00	93	94	96	97	97	99	99	100	100	100

From Table 1, we notice that the average weight of Population 1 does not seem to go below 0.7 for these scenarios. As n increases, the weight of Population 1 approaches 1, hence the MAMSE weights detect that the distributions are different and ultimately discard Population 2. Note that this convergence to 1 does not seem to occur for $\Delta = 0$ and seems very slow when Δ is tiny. The average weight for Population 1 increases as well when the discrepancy between the populations increases while n is kept fixed.

Table 2 shows the performance obtained for the MWLE with MAMSE weights when compared to the MLE. The ratio of the mean squared errors, 100 MSE(MLE)/MSE(MWLE) is shown; a value greater than 100 means that the MWLE is preferable. This ratio is akin to the relative efficiency of the MLE with respect to the MWLE.

TABLE 2: Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE). Samples of equal size n are simulated from normal distributions with unit variance and means 0 and Δ respectively. The results are averaged over 10000 replicates.

	Efficiency of the MWLE									
	$n = 5$	10	15	20	25	50	100	200	1000	10000
$\Delta = 0$	146	145	144	144	143	143	144	144	144	143
0.001	147	146	145	144	143	143	142	143	143	144
0.01	146	146	145	144	143	143	144	143	141	127
0.10	143	143	142	140	139	135	128	118	89	94
0.25	139	134	131	125	123	110	96	87	91	99
0.50	127	117	108	104	97	88	88	90	97	100
0.75	114	103	95	91	89	87	91	95	99	100
1.00	103	94	90	88	88	90	94	97	99	100
1.50	89	88	89	91	91	94	98	98	100	100
2.00	84	87	91	92	93	96	98	99	100	100

The MWLE performs better than the MLE for small n and Δ. When n and Δ increase, the two methods eventually perform equivalently. For the cases in between however, the MLE is a better choice than the MWLE. Fortunately, the loss (at most 16%) seems to be smaller than the potential gain (up to 47%). When the two populations are identical, a steady improvement of about 43% is observed. Note that we cannot expect to improve uniformly over the MLE since the mean is an admissible estimator.

The weighted likelihood could be especially useful in situations where a large population is available to support a few observations from the population of interest. For the next simulation, 40000 replicates of each scenario are produced with the same normal distributions as before, but with samples of size n and $10n$ for Population 1 and 2 respectively. Table 3 shows the average weight allocated to Population 1; Table 4 shows the relative efficiency of the methods as measured by 100 MSE(MLE)/MSE(MWLE).

TABLE 3: Average MAMSE weights for Population 1 when samples of size n and $10n$ are drawn from normal distributions with unit variance and means 0 and Δ respectively. The results are averages over 40000 replicates.

	Average values of $100\lambda_1$							
	$n=5$	10	15	20	25	50	100	200
$\Delta=0$	51	50	49	49	49	49	49	48
0.001	51	50	49	49	49	49	49	48
0.01	52	50	50	49	49	49	49	49
0.10	54	53	52	53	53	54	57	62
0.25	58	59	60	61	62	69	78	86
0.50	66	70	73	76	79	87	93	96
0.75	74	79	83	86	88	94	97	98
1.00	80	86	89	91	93	96	98	99
1.50	87	92	94	95	96	98	99	99
2.00	91	94	96	97	97	99	99	100

The general behavior of the weights is similar to that in the previous simulation, except that their minimal average value is below 0.5 this time around. As a consequence of its larger size, the sample from Population 2 gets a heavier weight.

It appears that a larger Population 2 magnifies the gains or losses observed previously. Fortunately however, the magnitude of the further improvements seem to exceed that of the extra losses.

Note that the MAMSE weights are invariant to a common transformation of the data in all populations. Therefore, simulation results would be identical (less simulation error) for normal populations with variance σ^2 and with means 0 and $\Delta\sigma$ respectively.

Overall, the MWLE works very well under the suggested scenarios.

6.2. Complementary populations.

We explained in Section 2 how the likelihood weights can be seen as mixing probabilities. Can the MAMSE weights detect and exploit the fact that Population 1 has the same distribution as a mixture of some of the other populations? Would the quality of the inference then be improved?

Pseudo-random samples of equal sizes n are drawn from the distributions

$$\text{Pop. } 1 : \mathsf{N}(0,1), \quad \text{Pop. } 2 : |\mathsf{N}(0,1)|, \quad \text{Pop. } 3 : -|\mathsf{N}(0,1)|$$

where $|\cdot|$ denotes absolute values. Hence Population 2 has a half-normal distribution and Population 3 follows the complementary distribution.

TABLE 4: Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE). Samples of sizes n and $10n$ are simulated from normal distributions with unit variance and means 0 and Δ respectively. The results are averaged over 40000 replicates.

	Efficiency of the MWLE							
	$n = 5$	10	15	20	25	50	100	200
$\Delta = 0$	223	223	223	222	222	221	222	221
0.001	223	225	223	221	222	223	221	220
0.01	223	222	222	220	221	221	220	218
0.10	216	209	203	197	191	169	142	113
0.25	187	165	147	135	125	100	83	78
0.50	139	111	97	90	85	79	83	89
0.75	111	91	85	82	82	85	90	94
1.00	98	85	84	83	85	90	94	97
1.50	88	86	88	89	90	94	97	98
2.00	86	89	91	92	93	96	98	99

We consider different sample sizes, each scenario being repeated 10000 times. The results are summarized in Table 5. The first column shows 100 MSE(MLE)/MSE(MWLE); the other columns show the average MAMSE weights allocated to each of the three populations.

First observe that the combined average MAMSE weight of Populations 2 and 3 accounts for at least half of the total weight for all sample sizes. The MAMSE weights thus detect that an equal mixture of Populations 2 and 3 share the same distribution as Population 1. Note also that the relative efficiency is uniformly greater than 100, which means that the MWLE with MAMSE weights is preferable to the MLE in these situations.

TABLE 5: Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE) and average MAMSE weights allocated to samples of sizes n drawn from $\mathsf{N}(0,1)$, $|\mathsf{N}(0,1)|$ and $-|\mathsf{N}(0,1)|$ respectively. The results are averages over 10000 repetitions.

n	Efficiency	$100\bar{\lambda}_1$	$100\bar{\lambda}_2$	$100\bar{\lambda}_3$
5	115	50	19	30
10	121	46	23	30
15	120	46	25	29
20	118	45	25	29
25	118	45	26	29
50	117	45	27	28
100	116	44	27	28
200	116	44	28	28
1000	115	44	28	28
10000	116	44	28	28

The column *Efficiency* shows 100 MSE (MLE)/MSEW (MWLE); the average MAMSE weights allocated to each of the three populations appears in the other columns.

6.3. Negative weights.

In most cases, the unconstrained optimization of $P(\boldsymbol{\lambda})$ yields positive weights. In some cases such as the one that we are going to explore, negative weights systematically occur. Some previous work such as van Eeden & Zidek (2004) showed that allowing negative weights may sometimes boost the performance of the MWLE. We explore the possibility of such improvements here.

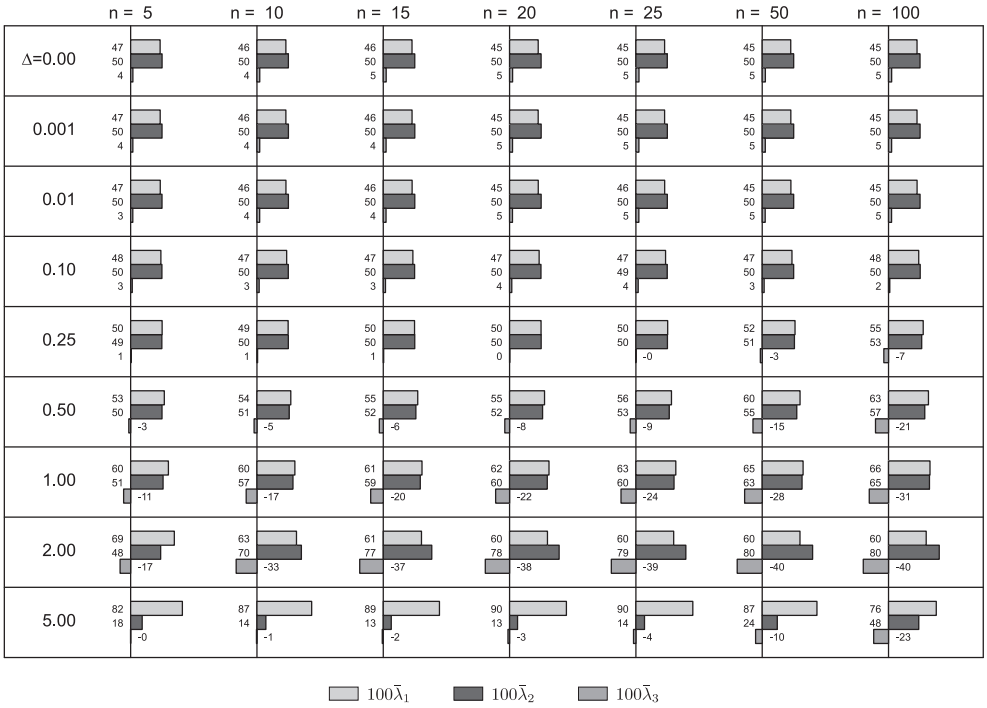

FIGURE 1: Average values of $100\times$ the MAMSE weights without the constraints $\lambda_i \geq 0$. Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathsf{N}(0,1)$ and a $\mathsf{N}(\Delta,1)$ distribution. All results are averages over 40000 repetitions.

Imagine a situation where a measurement of interest is cheaply obtained, but it is costly to determine whether a patient is diseased or not. We want to study the measurement of interest on the diseased patients. Suppose we have two small samples (one diseased, one not) as well as a larger sample where the health status of patients is unknown. If we allow negative values for MAMSE weights, would they adapt by including the larger population in the inference and allocating a negative weight to the small healthy population?

To represent the hypothetical situation above we simulate from the following distributions:

$$\text{Pop. } 1 : \mathsf{N}(0,1), \quad \text{Pop. } 2 : 0.5\mathsf{N}(0,1) + 0.5\mathsf{N}(\Delta,1), \quad \text{Pop. } 3 : \mathsf{N}(\Delta,1),$$

where Population 1 and 3 have equal sample sizes of n, but Population 2 has a sample size of $10n$. Each scenario is repeated 40000 times.

Although we allow weights to be negative, we still apply the preprocessing step and set the weight of a population to 0 when it does not overlap with the sample from Population 1. If the preprocessing were ignored, a nonnegative definite $\bar{\mathsf{A}}$ could occur occasionally, and then the MAMSE weights would not be unique.

Applying the preprocessing does not affect the pertinence of this example: if the distributions in the populations of diseased and healthy are so different that the samples are often disjoint, there is no point in using the weighted likelihood to include Population 2 as the measurements

are in fact a cheap diagnostic test. Moreover, previous simulations without preprocessing yielded results that are not better than those presented here.

Figure 1 shows the average values of the unconstrained MAMSE weights for different scenarios. Negative weights do appear, hence the MAMSE criterion detects that Population 2 is a mixture of the other two populations and removes the component which is not of interest.

For a large Δ, notice how the negative weights are closer to 0 for smaller samples. In such cases, there is a higher probability that the sample from Population 3 will be disjoint of the sample from Population 1. As a result, the weight allocated to Population 3 is more often forced to 0 by the preprocessing step. As the sample sizes increase, the samples overlap more frequently.

Table 6 shows the performances obtained by the MWLE with unconstrained MAMSE weights. The MWLE performs better than the MLE in most cases, being almost twice as good in many cases. Unfortunately, the performances for large Δ are very poor, especially in the cases where the difference between the populations is so large that they overlap only slightly.

TABLE 6: Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE) when the MAMSE weights are calculated without the constraints $\lambda_i \geq 0$. Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathsf{N}(0,1)$ and a $\mathsf{N}(\Delta,1)$ distribution. All results are averages over 40000 repetitions.

	100 MSE(MLE)/MSE(MWLE)						
	$n=5$	10	15	20	25	50	100
$\Delta=0$	195	196	197	198	197	197	198
0.001	196	196	197	197	198	198	197
0.01	196	196	197	197	198	198	197
0.10	195	194	194	194	192	184	172
0.25	190	182	176	170	165	144	121
0.50	173	153	140	131	124	107	97
1.00	137	113	105	101	100	97	96
2.00	116	92	86	84	84	84	84
5.00	51	49	51	54	57	62	55

Using a weighted likelihood with negative weights provides an improvement over the MLE, but a similar improvement may be obtainable when the constraints are enforced. Table 7 shows the performance of the MWLE when the usual MAMSE weights are used. Figure 2 shows the average values of the weights obtained in that case. Using the MWLE with positively constrained MAMSE weights also provides an improvement over the MLE. This improvement is sometimes larger than that obtained with unconstrained weights. To discern between the two versions of MAMSE weights, Table 8 compares their relative efficiency; values above 100 favor the unconstrained weights. Note that the standard deviation of the error due to simulation in Table 8 can be more than one unit, but does not exceed 1.3 units.

It seems that allowing negative weights further improves the performances only in a few cases. In fact, Figure 2 shows that Population 2 by itself can be used and Table 7 shows it has a positive impact. Table 8 suggests that the constrained MAMSE weights are to be preferred more often than not. If we consider other complications that arise from allowing negative weights, (e.g., making the weighted empirical distribution function nonmonotone) keeping the constraints $\lambda_i \geq 0$ in the definition of the MAMSE weights seems a better option.

A different prevalence of the diseased in Population 2 could affect the simulation results. If major differences were observed, the conclusion above could be revisited.

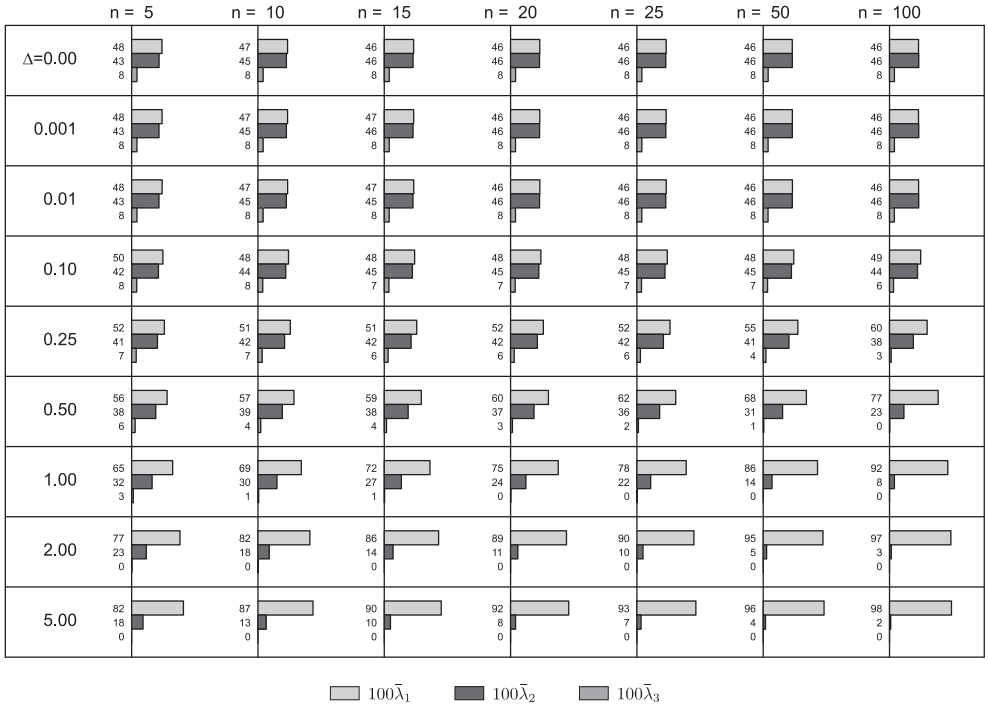

FIGURE 2: Average values of $100\times$ the usual MAMSE weights (with constraints $\lambda_i \geq 0$). Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathsf{N}(0,1)$ and a $\mathsf{N}(\Delta,1)$ distribution. All results are averages over 40000 repetitions.

TABLE 7: Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE) when the usual MAMSE weights (i.e., constrained to positive values) are used. Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathsf{N}(0,1)$ and a $\mathsf{N}(\Delta,1)$ distribution. All results are averages over 40000 repetitions.

	100 MSE(MLE)/MSE(MWLE)						
	$n=5$	10	15	20	25	50	100
$\Delta=0$	211	209	210	210	209	208	208
0.001	212	210	209	209	210	209	208
0.01	212	210	210	209	210	209	208
0.10	212	209	207	206	203	194	180
0.25	207	196	187	180	173	146	118
0.50	186	161	144	131	122	98	82
1.00	139	111	97	89	86	79	82
2.00	97	82	79	78	79	84	90
5.00	51	48	50	53	57	68	79

TABLE 8: Relative efficiency of the MWLE with and without the constraints $\lambda_i \geq 0$ as measured by 100 MSE(constrained MWLE)/MSE(unconstrained MWLE). Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathsf{N}(0,1)$ and a $\mathsf{N}(\Delta,1)$ distribution. All results are averages over 40000 repetitions.

	100 MSE(constrained)/MSE(negative)						
	$n=5$	10	15	20	25	50	100
$\Delta = 0$	92	94	94	94	95	95	95
0.001	92	93	94	94	94	95	95
0.01	92	93	94	94	94	95	95
0.10	92	93	94	94	94	95	96
0.25	92	93	94	95	96	98	102
0.50	93	95	98	100	102	109	119
1.00	99	102	109	114	117	123	117
2.00	119	112	109	107	107	100	94
5.00	100	101	102	102	101	91	69

6.4. Earthquake data.

We now use a model whose weighted likelihood estimate does not have a simple form, i.e., it is not a weighted average of the MLE of each population.

Natural Resources Canada http://earthquakescanada.nrcan.gc.ca/ maintains an educational website with resources about earthquakes. From their website, it is possible to download data about recent western Canadian earthquakes. The histograms in Figure 3 show the magnitude of the earthquakes that occurred in the 5-year period from 12 February 2001 to 12 February 2006. Events are divided into 3 groups depending on the geographical location of their epicenter. For the purpose of this example, we make the assumption that the magnitudes of the earthquakes are independent random variables and fit a gamma distribution to each of the three populations using maximum likelihood. The fitted curves appear on Figure 3 and the estimated values of their parameters are shown in Table 9 along with the number of observations in each area. The gamma model is parametrized as

$$f(x \,|\, \beta, \mu) = \frac{\beta^{\beta\mu}}{\Gamma(\beta\mu)} x^{\beta\mu - 1} e^{-\beta x}$$

for $\beta, \mu, x > 0$.

TABLE 9: Number of earthquakes in three areas of western Canada between 12 February 2001 and 12 February 2006. The magnitude of these earthquakes is modeled by a gamma distribution; the maximum likelihood estimates appear below and are used as the "true" parameters for this simulation.

	Lower Mainland – Vancouver Island	Elsewhere in BC or in Alberta	Yukon and North West Territories
β	1.654	2.357	6.806
μ	1.437	1.869	2.782
n	4743	4866	1621

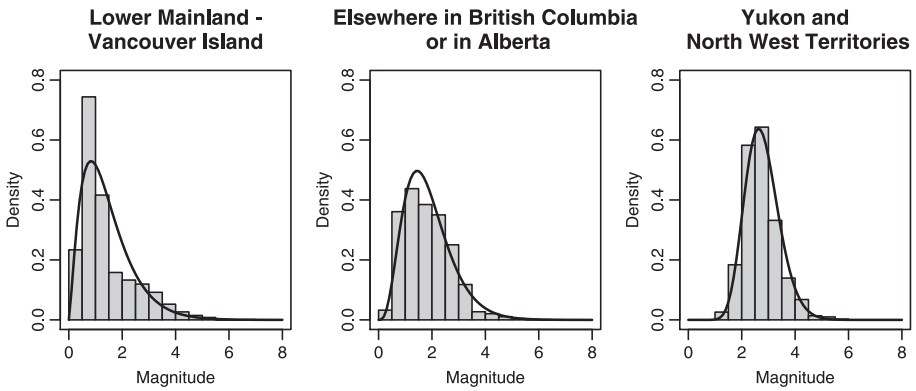

FIGURE 3: Histograms of the magnitude of earthquakes measured between 12 February 2001 and 12 February 2006 for three different areas of western Canada. The curves correspond to the fitted gamma density.

We focus our interest on the magnitude of the next earthquake with epicenter in the Lower Mainland – Vancouver Island area. Suppose that only the 50 most recent events from each of the three regions are available. Would the MWLE that uses data from all three regions provide a better estimate than the MLE? To investigate the question, we produce 10000 pseudo-random samples of earthquakes based on the fitted gamma models shown above.

The average MAMSE weights are 0.959 for the Lower Mainland – Vancouver Island area, 0.041 for the rest of British Columbia and Alberta and finally, nearly 0 for Yukon and North West Territories. Although it looks like a small contribution, the MSE of the MWLE for the vector (β, μ) was smaller with 100 MSE(MLE)/MSE(MWLE)=107.

We also considered other values of possible interest, namely some probabilities about the magnitude (M) of the next earthquake that are all obtained by plugging the MLE or MWLE in the gamma model. Table 10 summarizes these results.

TABLE 10: Efficiency in estimating some probabilities about the magnitude of the next earthquake in the Lower Mainland – Vancouver Island area followed by the average of the actual estimates and their true values. Efficiency is measured by 100 MSE(plug-in MLE)/MSE(plug-in MWLE). The following four columns contain different probabilities that must be multiplied by the corresponding multiplier.

Prob	Efficiency	MLE	MWLE	Model	Data	Multiplier
$\mathrm{P}(M > 1)$	123	62	63	68	51	$\times 10^{-2}$
$\mathrm{P}(M > 2)$	114	22	24	40	22	$\times 10^{-2}$
$\mathrm{P}(M > 3)$	112	66	73	174	98	$\times 10^{-3}$
$\mathrm{P}(M > 4)$	113	19	21	51	26	$\times 10^{-3}$
$\mathrm{P}(M > 5)$	112	51	59	99	53	$\times 10^{-4}$
$\mathrm{P}(M > 6)$	80	14	17	12	6	$\times 10^{-4}$

The column *Efficiency* of Table 10 corresponds to the relative efficiency of using the MWLE compared to using the MLE as plug-in parameters for the gamma model in order to evaluate the probability of interest. The numbers shown are 100 MSE(plug-in MLE)/MSE(plug-in MWLE) followed by the estimated values of $P(M > k)$ using the MLE and the MWLE as plug-in parameters. For comparison purposes, the columns *Model* and *Data* contain respectively the true probabilities (from the simulated model) and the empirical proportions in the complete data set. All probabilities are scaled for easier reading; using the corresponding multiplier will yield the

original value. Note that discrepancies with the empirical probabilities reveal weaknesses of the gamma model to perfectly represent the magnitude of earthquakes rather than an advantage for one method over the other.

Interestingly enough, the MSE of the estimates is almost always smaller with the MWLE. Improved performance is hence possible by using the MWLE with MAMSE weights in this situation with distributions copied from real life.

7. ASYMPTOTIC PROPERTIES

Because they are calculated from the data, the MAMSE weights are random variables. Hu (1997) proves the weak consistency and the asymptotic normality of the maximum weighted likelihood estimate, but his results hold only for fixed weights, i.e., weights that may depend on sample sizes, but that are not random variables.

For the case of adaptive weights such as the MAMSE weights, further work has been done by Wang, van Eeden & Zidek (2004). They prove the consistency and normality of the maximum weighted likelihood estimate under the assumption that the weights shift entirely to the population of interest, i.e., $\boldsymbol{\lambda} \to [1, 0, \ldots, 0]^{\mathsf{T}}$, at a specified rate as the sample sizes of all populations go to infinity. The simulations of Section 6 seem to indicate that the MAMSE weights do not behave that way. When a mixture of the additional populations is identical to the target, the weights are shared between these populations even for very large sample sizes.

The MAMSE weights minimize $P(\boldsymbol{\lambda})$ and hence guarantee that $\boldsymbol{\lambda}' = [1, 0, \ldots, 0]^{\mathsf{T}}$ is a suboptimal choice, which implies that

$$\begin{aligned}\int \left\{\widehat{F}_1(x) - \widehat{F}_{\boldsymbol{\lambda}}(x)\right\}^2 \mathrm{d}\widehat{F}_1(x) \;&\leq\; \int \left[\{\widehat{F}_1(x) - \widehat{F}_{\boldsymbol{\lambda}}(x)\}^2 + \widehat{\mathrm{var}}\{\widehat{F}_{\boldsymbol{\lambda}}(x)\}\right] \mathrm{d}\widehat{F}_1(x) \\ &=\; P(\boldsymbol{\lambda}) \leq P(\boldsymbol{\lambda}') = \int \frac{1}{n_1}\widehat{F}_1(x)\{1 - \widehat{F}_1(x)\}\,\mathrm{d}\widehat{F}_1(x) \leq \frac{1}{4n_1}.\end{aligned}$$

Therefore, as the sample size from Population 1 increases, the mixture of empirical distributions $\widehat{F}_{\boldsymbol{\lambda}}$ must become very close to $\widehat{F}_1$ which is known to converge uniformly and almost surely to the target distribution F_1. Recall the heuristic development of Section 2: the MWLE maximizes the proximity between $F(x \,|\, \theta)$ and $\widehat{F}_{\boldsymbol{\lambda}}$.

Asymptotic properties of $\widehat{F}_{\boldsymbol{\lambda}}$ and of the MWLE are developed in Plante (2007). The proofs require a detailed treatment since standard convergence results do not apply to the MAMSE weights. In all cases, the heuristic argument above indicates why the MAMSE weights have good asymptotic properties, despite the fact that we do not assume the proximity of the m populations.

8. CONCLUSION

The weighted likelihood is a method that allows one to include relevant information from available data even if they do not exactly follow the target distribution. The paradigm that we use throughout this paper has been around for a few years now, but the absence of an efficient and reliable method for determining likelihood weights undoubtedly limited its popularity.

In this paper, we suggest a reliable nonparametric method for determining adaptive weights and we provide an algorithm for calculating them. We then show through simulations that the MWLE using MAMSE weights often performs better than the MLE. These good performances hold in an example where the simulated models mimic distributions based on real data.

Plante (2007) studies the asymptotic properties of the MAMSE weights as well as their extension to multivariate data and censored data. More work could be done in these directions.

The original work of Hu (1994) is based on a paradigm inspired by smoothing problems where each datum may have a different weight. Revisiting this paradigm with the heuristic of Section 2 and the idea of MAMSE weights could be fruitful, especially if we try to link it to useful applications. For instance, Hu & Rosenberg (2000) use such a weighted likelihood to make

inferences about a process that reaches stability after a certain number of iterations. Their work is in the same spirit as Hu, Rosenberg & Zidek (2000) where the weighted likelihood is used to make inferences about dependent data. To extend the MAMSE weights to such situations, we could create subgroups of the data and see them as populations. Another approach could consist in using parametric models to infer the cumulative distribution function in the MAMSE criterion rather than their empirical counterparts. Such extensions are however left to future work. Meanwhile, we hope that the MAMSE weights will contribute to popularizing the weighted likelihood so that analysts may take advantage of its ability to borrow strength with minimal assumptions.

ACKNOWLEDGEMENTS

This paper constitutes part of the author's doctoral thesis supervised by Professor James V. Zidek at the University of British Columbia. I am thankful to Professor Zidek for his guidance and support in discovering the weighted likelihood and in developing the results that led to this publication. For partial support of this work through graduate scholarships and research grants, thanks are due to the Natural Sciences and Engineering Research Council of Canada and to the Fonds québécois de la recherche sur la nature et les technologies.

REFERENCES

H. Akaike (1977). On entropy maximization principle. In *Applications of Statistics: Proceedings of the symposium held at Wright State University, Dayton, Ohio, 14–18 June 1976* (P. R. Krishnaiah, ed.), North Holland, Amsterdam, pp. 27–42.

B. Efron (1996). Empirical Bayes methods for combining likelihoods. *Journal of the American Statistical Association*, 91, 538–550.

F. Hu (1994). *Relevance Weighted Smoothing and a New Bootstrap Method*. Unpublished doctoral dissertation, Department of Statistics, The University of British Columbia, 177 pp.

F. Hu (1997). The asymptotic properties of the maximum-relevance weighted likelihood estimators. *The Canadian Journal of Statistics*, 25, 45–59.

F. Hu & W. F. Rosenberg (2000). Analysis of time trends in adaptive designs with application to a neurophysiology experiment. *Statistics in Medicine*, 19, 2067–2075.

F. Hu, W. F. Rosenberg & J. V. Zidek (2000). Relevance weighted likelihood for dependent data. *Metrika*, 51, 223–243.

F. Hu and J. V. Zidek (1993). *A relevance weighted nonparametric quantile estimator*. Technical report no. 134, Department of Statistics, The University of British Columbia, Vancouver.

F. Hu and J. V. Zidek (2002). The weighted likelihood. *The Canadian Journal of Statistics*, 30, 347–371.

D. G. Luenberger (2003). *Linear and Nonlinear Programming*, Second edition. Kluwer, Boston.

J.-F. Plante (2007). *Adaptive Likelihood Weights and Mixtures of Empirical Distributions*. Unpublished doctoral dissertation, Department of Statistics, The University of British Columbia, Vancouver, 171 pp.

C. Stein (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proceedings of the Third Berkeley Symposium Mathematical Statistics and Probability*, 1, 197–206.

C. van Eeden & J. V. Zidek (2004). Combining the data from two normal populations to estimate the mean of one when their means difference is bounded. *Journal of Multivariate Analysis*, 88, 19–46.

X. Wang (2001). *Maximum Weighted Likelihood Estimation*. Unpublished doctoral dissertation, Department of Statistics, The University of British Columbia, Vancouver, 151 pp.

X. Wang, C. van Eeden & J. V. Zidek (2004). Asymptotic properties of maximum weighted likelihood estimators. *Journal of Statistical Planning and Inference*, 119, 37–54.

X. Wang & J. V. Zidek (2005). Selecting likelihood weights by cross-validation. *The Annals of Statistics*, 33, 463–501.

Received 6 August 2007
Accepted 26 March 2008

Jean-François PLANTE: plante@utstat.toronto.edu
Department of Statistics, University of Toronto
Toronto, Ontario, Canada M5S 3G3

SSC Statistical Society of Canada
Société Statistique du Canada